Environmental Health Perspectives Supplements
Vol. 102 Suppl 1:33-38 1994

Multiple Comparisons in Long-Term Toxicity Studies

by Ludwig Hothorn

Several multiple comparison procedures (MCPs) are discussed in relation to the specific formulation of type I and type II errors in toxicity studies and the typical one-way design control versus *k* treatment/dose groups. Examples of these MCPs are: the standard many-to-one MCP (Dunnett's procedure), sequential rejection modifications, closed testing procedures, many-to-one MCPs with an ordered alternative hypothesis, procedures based on the assumption of a mixing distribution of responders and nonresponders, and MCP's for multiple end points.

Introduction

Why is it that multiple comparison procedures (MCPs) are being discussed in toxicology even today, despite the fact that they are every-day procedures in biostatistics? This paper deals with several sources of multiplicity in long-term toxicity studies and possible methods for suitable statistical analysis.

Based on the closed testing principle discussed by Marcus et al. (*1*), a revolution in MCPs has taken place. We can thus diminish the antagonism enforcing α_{exp} (type I error) and decreasing the power π (where $\pi = 1 - \beta$, β ... type II error). This paper presents a special case where α_{exp} is held and the maximum power of the two-sample case is guaranteed. This paper is therefore limited to regulatory toxicity studies, e.g., carcinogenicity, mutagenicity, according to national/international guidelines, for example, the European Community (EC) guideline (*2*). Regulatory toxicity studies are so-called safety studies, the purpose of which is to ascertain carcinogenic, mutagenic side effects etc. For this purpose, the statistical hypothesis in relation to type I and II errors should be specified: *a*) The risk of a type I error, α, represents the producer's risk: the conclusion is therefore that a toxic side effect exists, while in fact this is not the case. *b*) The risk of a type II error, β, represents the customer's risk: the conclusion is therefore that a toxic effect does not exist, while in truth one actually does. Intuitively, it is clear that both risks must be handled with care, even though controlling the type II error should be of primary concern in toxicology.

Usually, the type II error is defined comparisonwise and the type I error experimentwise (α_{exp}). A typical design analyses comparisons between the control and treatment/dose groups, several time points, both sexes, elements of a multivariate end point vector, and multiple tumor sites. Because of a dramatic increase in the type II error with such a high-dimensional design, an α_{exp} formulation is normally used for the subdesign control versus k treatment/dose groups. The purpose of an adequate statistical analysis is to minimize the type II error while holding α_{exp} constant. This article will therefore investigate several MCPs to establish the conditions under which the above requirement can be fulfilled.

Experimental Design of Long-Term Toxicity Studies

In long-term toxicity studies, there are three types of experimental design that can be distinguished for the above-mentioned subdesign. *a*) Control, dose$_1$, . . . , dose$_k$, where $C = 0 < D_1 < \ldots < D_k$, the purpose of which is to analyze dose response analysis or estimate the no-observed-effect dose. *b*) Control, treatment$_1$ . . . , treatment$_k$, with treatment T_j . . . several substances, combinations, etc. The purpose is to characterize all contrasts {control versus T_j $\forall$ $j \epsilon$ (1, . . . , k)} *c*) Control, {D_j or T_j}, P^+. The purpose of using a positive control group, P^+ (administration of a known toxic substance), is to check the sensitivity of the test system currently in use (animals, bacteria, etc.). Using this simple closed testing procedure, α_{exp} can also be held constant in this most complex design (L. Hothorn, in preparation).

Multiple Comparison versus Modeling

Two widely used and disjointed types of statistical approach are possible for long-term toxicity studies: modeling, choosing a suitable dose–response model and fitting the model to the data, e.g., for the AMES assay according to Margolin et al. (*3*); and MCPs. This paper only discusses MCPs.

MCPs are suitable for all three above-mentioned types of design. Modeling is sometimes uncertain for the typical guideline-related two or three dose-groups design. MCPs usually use fewer *a priori* assumptions (e.g., no problems with a correct model choice). An interaction of incorrect model choice and estimation error in the modeling approach is possible. Robust use

Department of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, D-6900 Heidelberg, Germany.

This paper was presented at the International Biostatistics Conference on the Study of Toxicology that was held May 13–25, 1991, in Tokyo, Japan.

of MCPs in routine evaluation of studies with multiple end points is possible. Of course, the MCP approach also has several disadvantages, such as no possibility of extrapolation.

MCPs in Control versus *k* Treatment/ Dose Groups Design

Two-Sample versus *k*-Sample Testing

Toxicology journals often contain papers in which the statistical analysis is based on the two-sample *t*-test or the Wilcoxon-Mann-Whitney *U* test, even in the *k*-sample many-to-one situation (*4*) using a comparisonwise α_{comp} level; i.e., testing each contrast with α for example on a 0.05 level independently. Using this simple approach, the experimentwise α_{exp}-level is violated on the one hand, whereas, on the other, the type II error is smaller in comparison with an MCP and does not depend on the number of treatment/dose groups. This is the testing dilemma always faced in toxicity studies. Several compromises and an ideal situation (minimum type II error and holding α_{exp}) will now be discussed.

Many-to-one MCPs can be recommended on the whole. But if two-sample tests are used, then they should be used only for the contrasts $(C - D_j)$, but not for the between-dose contrasts, $(D_j - D_i)$ with $(i \neq j) \in (1, \ldots, k)$.

k-Sample Tests versus *k*-Sample Procedures

There is some desire to clarify the difference between tests and MCPs from both a toxicological and a biostatistical viewpoint. A *k*-sample test, e.g., the well-known *F*-test, represents a single decision problem:

$$\mathbf{H_0 : F_C = F_{D_1} = ... = F_{D_k}}$$

$$\mathbf{H_A : F_C \neq F_{D_1} \neq ... \neq F_{D_k}}$$

with *F* distribution function for testing the global substance effect. An MCP represents a multiple decision problem:

$$\mathbf{H_A^{ij} : F_i \neq F_j \ \forall \ (i \neq j) \in (1 \ ..., k)}$$

for testing every contrast $(C - D_j)\ \forall j \in (1, \ldots, k)$. Because not only the global effect, but also each single contrast $(C - D_j)$ is of interest in toxicology, application of MCP is recommended. A combination of both approaches based on the closed testing principle is also possible, providing both global and local information.

All-Pair versus Many-to-One Procedures

Commonly used statistical software packages are generally oriented to all-pair MCPs, such as Tukey, Scheffe, Duncan, etc. All-pair MCPs analyze not only contrasts of interests $(C - D_j)$ but also contrasts $(D_j - D_i)$ with $(i \neq j) \in (1, \ldots, k)$. The type II error rate thus increases (*5*): control versus k=3 dose groups, $\alpha_{exp} = 0.05$, $\sigma/d = 1.0$; $n_j = 24$ (with σ end point-specific variance, *d* detectable difference); many-to-one MCP (Dunnett) $\beta = 0.106$; all-pair MCP (Tukey) $\beta = 0.200$.

The Standard Many-to-One MCP: Dunnett's Procedure

In control versus *k* treatment design, Dunnett's (*6*) procedure is commonly used to approximate normally distributed end points. Other types of end points occurring in toxicology will not be discussed in this paper. For dichotomous end points, see Piegorsch (*7*).

Hypothesis formulation:

$$\mathbf{H_0 : \mu_C = \mu_{T_j} \ \forall \ j \in (1, ..., k)}$$

$$\mathbf{H_A^j : \mu_C < \mu_{T_j}}$$

without limitation, for a one-sided increase, with μ_j expected value.

Test statistics: $\forall \ \mathrm{j} \in (1, ..., \mathrm{k}):$

$$\mathrm{d_j} = (\bar{\mathrm{x}}_\mathrm{j} - \bar{\mathrm{x}}_\mathrm{C})/\sqrt{\mathrm{MQ_R}(1/\mathrm{n_C} + 1/\mathrm{n_j})}$$

with MQ_R the mean-square-error estimator.

Decision rule: H_0 will be rejected if:

$$\mathrm{d_j} > \mathrm{d_{k,df,c_j,1-\alpha,one-sided}}$$

with $$\mathrm{df} = \textstyle\sum_{\mathrm{j=C}}^{\mathrm{k}}(\mathrm{n_j} - 1) \quad \mathrm{j} \in (\mathrm{C}, 1, ..., \mathrm{k})$$

$$\mathrm{c_j} = 1/\sqrt{\mathrm{n_C}/\mathrm{n_j} + 1}$$

The quantiles $d_{k,\ df,\ c_j,\ 1\ -\ \alpha,\ \mathrm{two\ /\ one-sided}}$ are available from tables (*8–10*) or computer programs are available for calculation (*11*).

Dunnett's procedure is relatively robust against violation of the normal distribution assumption (*12*, Ortseifen and Hothorn, in preparation). However, for $n_j > 10$, the nonparametric analog according to Steel (*13*) shows a better power behavior (even in the near normal distributed case).

The maximum power of Dunnett's procedure is attained with: $n_c = \sqrt{k}n_j$ (*14–16*). This is not the case for the Williams (*17*) procedure, assuming an ordered alternative (*18*).

The power depends on the number of treatment or dose groups *k*, which implies that inclusion of further nonsignificant treatment groups can lead to overlooking significant effects (*19*). A rule for design using MCPs is to use only the minimal necessary number of treatment/dose groups.

In the case of variance heterogenicity, Dunnett's procedure is not robust (*12*). Other approaches should be used in this case, e.g., α-adjusted Welch-tests or Brownie (*20*) type of control group variance inclusion (Hothorn and Ortseifen, in preparation).

Simultaneous versus Sequential Rejective Procedures

The closed testing principle in many-to-one MCPs is quite simple (in comparison with all-pair MCPs) because a complete system of hypotheses with $(2^k - 1)$ elementary hypotheses (*21*) is given. Several types of sequential rejective modifications will be discussed: *a*) Bonferroni/Holm (*22*) procedure based on two-sample tests, *b*) sequential rejection modification according to Marcus et al. (*1*) or Hayter and Tamhane (*23*), *c*) Hommel (*24*, *25*)/Hochberg (*26*) / Rom (*27*) reverse Holm procedure based on two-sample tests, *d*) closed testing procedure based on global tests (*5*), *e*) procedure with *a priori* hierarchical hypotheses (*28*).

Bonferroni/Holm Procedure. Use specific two-sample tests for the elementary contrasts, $(C - D_j)$, Order the related *p*-values:

$$p_j : p_{(1)} \leq p_{(2)} \leq, \ldots, \leq p_{(k)}$$

Decision scheme: if

$$p_{(1)} > \alpha/k =\pm> \text{STOP } H_0^{(1)}, \ldots, H_0^{(k)}$$

cannot be rejected, otherwise go to the next step: if

$$p_{(2)} > \alpha/(k-1) ==> \text{STOP } H_A^{(1)}\ H_0^{(2)}, \ldots, H_0^{(k)}$$

is valid and $H_0^{(2)} \ldots, H_0^{(k)}$ cannot be rejected, etc.

Marcus et al. Modification. Use the Dunnett statistics:

$$\forall\, j \in (1, \ldots, k):$$

$$d_j = (\bar{x}_j - \bar{x}_C)/\sqrt{MQ_R(1/n_C + 1/n_j)}$$

Order the test statistics:

$$d_{(1)} \leq d_{(2)} \leq, \ldots, \leq d_{(k)}$$

Decision scheme: if

$$d_{(k)} < d_{k,df,c_j,1-\alpha,\text{one-sided}} ==> \text{STOP } H_0^{(1)}, \ldots, H_0^{(k)}$$

cannot be rejected, otherwise go to the next step.

$$d_{(k-1)} < d_{k-1,df,c_j,1-\alpha,\text{one-sided}} ==> \text{STOP } H_A^{(k)}\ H_0^{(k-1)}$$

is valid and $H_0^{(1)}, \ldots, H_0^{(K-1)}$ cannot be rejected, etc.

Hochberg Modification. The Hochberg modification is the numerically simplest version. Use specific two-sample tests for the elementary contrasts, $(C - D_j)$, Order the related *p*-values:

$$p_j : p_{(1)} \leq p_{(2)} \leq, \ldots, \leq p_{(k)}$$

Decision scheme: if

$$p_{(k)} < \alpha ==> \text{all } H_0^{(1)}, \ldots, H_0^{(k)}$$

are rejected and STOP, otherwise $H_0^{(k)}$is valid and go to the next step. If

$$p_{(k-1)} < \alpha/2 ==> \text{all } H_0^{(1)}, \ldots, H_0^{(k-1)} - H_0^{(k-1)}$$

are rejected and STOP, otherwise $H_0^{(k-l)}$is valid and go to the next step, etc.

The important difference between these three modifications is that the Marcus et al. modification is based on an MCP and causes a dimension reduction in *k*, while the others are based on two-sample tests and cause an α reduction

Holm Modification and Closed Testing Procedure. For the Holm modification: the first step is p_{min} versus α/k, in contrast to the Hochberg modification, where the first step is p_{max} versus α. The Holm modification is more powerful.

The closed testing procedure is based on a global test. Use suitable *j*-dimensional many-to-one test statistics $j \in (1, \ldots, k)$, e.g., Fligner/Wolfe contrast test (*29*). The testing strategy (for simplicity, given here as *C*, $k = 3$) is shown in Figure 1. This multiple procedure works simply as follows: A level α test is performed on stage 1. If and only if the $H_0^{\text{stage 1}}$ is rejected, all subhypotheses at stage 2 are tested on the same α level, and so on. If a $H_0^{(\ldots)}$ is not rejected, none of the subhypotheses are rejected.

Procedure with* a priori *Hierarchical Hypotheses. Use specific two-sample tests for the elementary contrasts, $(C - D_j)$, and estimate the related *p*-values (without ordering). Decision scheme: if

$$p_k > \alpha ==> \text{STOP } H_0^1, \ldots, H_0^k$$

cannot be rejected, otherwise H_0^k is rejected, and go to the dose level (*k*-1). If

$$p_{k-1} > \alpha ==> \text{STOP } H_0^1, \ldots H_0^{k-1}$$

cannot be rejected, otherwise $H_0^{(k-1)}$ is rejected, and go to the dose level (*k*-2), etc.

This procedure represents a special case of the closed testing procedure under the assumption of an ordered alternative hypothesis. If the p_j values within a real study are ordered, then with this procedure we find an ideal situation in MCP: holding

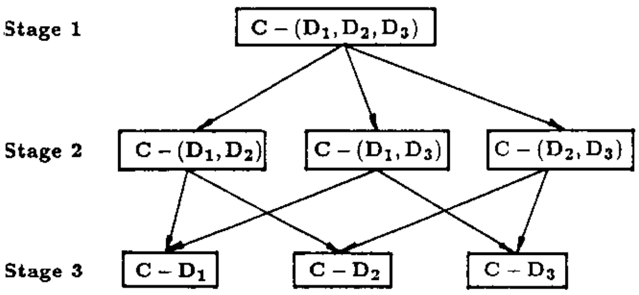

FIGURE 1: Complete hypotheses system in the case of a control and three dose groups.

α_{exp} and guaranteeing the maximum powers π_j of the two-sample tests (based on comparisonwise α). This procedure is moderately robust against violations of this monotonicity assumption (*28*).

Nonrestricted versus Ordered Alternative Hypotheses

Now we will consider the design $C, D_1, \ldots, D_k$. Assuming a monotonic dependence of the effect on dose, restriction of alternative hypotheses is possible:

$$\begin{gathered} H_A : F_C \leq F_{D_1} \leq \ldots \leq F_{D_k} \\ \text{at least } F_C < F_{D_k} \end{gathered}$$

With this restriction, an increase in power in relation to the MCPs with unrestricted alternative hypotheses can be expected. Possible MCPs are *a*) simultaneous MCPs: for continuous, normally distributed end points, the analogue of Dunnett's MCP is the Williams (*17,30*) procedure. For the nonparametric case, the analogue of Steel's MCP is the Shirley (*31,32*) procedure. *b*) sequential rejection MCPs: MCP on *a priori* ordered hypotheses, based on any two-sample tests. For binomially distributed end points, the closed testing procedure is based on Armitage's (*33*) trend test (*19*). For Poisson-distributed end points, the closed testing procedure is based on Lee's (*34*) trend test (*19*).

Comparison of Several Procedures with Simulation Studies

For commonly observed conditions of real toxicity study data, namely expected value profiles, dimension of *k*, sample sizes n_j, α levels, variances, etc., several procedures were investigated with simulation studies (*5,18,28,30, 35–37*). For practical application, these simulation results can be summarized in a rather simple way: recommendation of the Hommel (*24*) / Hochberg (*26*) procedure, without restriction of the alternative hypothesis, a power behavior near the MCPs with ordered alternative was observed. It should be pointed out that for sequential rejection procedures, the estimation of confidence intervals in time was not solved satisfactorily.

Unimodal versus Mixing Distribution Assumption

All MCPs discussed in the preceding sections compare expected values. In real data, two situations may occur: greater variability (variance) with increasing response and existence of a subpopulation of nonresponders. This problem can be treated by several approaches: *a*) use of MCPs that are robust under variance heterogenicity (Hothorn and Ortseifen in preparation); *b*) so-called location-scale models, e.g., a combination of the U-test (location) and Ansari/Bradley (*38*) test [scale (*39*)] or the Brownie (*20*) type of control group variance includion; *c*) assumption of a mixing distribution of responders and nonresponders with the following hypotheses:

$$H_0 : F_C(x) = F_D(x)$$

$$\begin{aligned} H_A : F_C(x) < F_D(x) \text{ with } F_D(x) \\ = (1 - r)F_C(x) + r\,F_{patho}(x) \end{aligned}$$

where *r* is unknown, $(1-r)$ is the proportion of nonresponders, and *r* is the proportion of responders.

Two types of Lehmann (*40*) alternative will be considered here: shift:

$$F_{patho}(x) = F_C(x - \delta)$$

according to Good (*41*) and power:

$$F_{patho}(x) = F_C^a(x)$$

according to Lehmann (*40*). Johnson et al. (*42*) suggested, for the shift alternative, approximate score statistics based on following mixed normal score function:

$$\begin{aligned} &sm(i) \\ &= \exp(-d^2/2)\exp(d\Phi^{-1}(i/(n_C + n_T + 1))^{-1} \end{aligned}$$

where *i* is a rank in the combined (control+treatment) sample, *d* is a constant (in the simulation study where *d*=0.5,1,1.5,2 were used; only the case *d*=1 will be reported here), and Φ^{-1} is a distribution function of the normal distribution.

As a generalization of Wilcoxon-Mann-Whitney (WMW) scores, Conover and Salsburg (*43*) proposed the following approximate score function for the power alternative:

$$sc(i) = (i/(n_C + n_T + 1))^{a-1}$$

where *a* is an integer constant (*a*=3,4,5,6 were used in the simulation study; here, only the case *a*=4 will be reported).

In toxicology, tests based on this mixing distribution assumption were used for behavioral studies (*44*), teratological studies (*45*), sister chromatid exchange mutagenicity assays (*42*), chronic studies (*5*), and micronucleus mutagenicity assays (*46*). With simulation studies (*42, 46*), advantages in power can be shown for several practical data situations in toxicology.

Many-to-One MCPs for Multiple End Points

In long-term toxicity studies, several end points occur, (*19*): approximate, normally distributed (e.g., body mass); nonnormally distributed [e.g., the skewed distributed liver enzyme ASAT (*5*)]; binomially distributed (e.g., tumor rate); Poisson-distributed (e.g., number of tumors). The commonly used evaluation consists of separate univariate analysis of each single end point, e.g., Unkelbach et al. (*47*), but a multivariate analysis of multiple end points in the many-to-one design is also possible: *a*) T^2 modification according to Higazi and Dayton (*48*), *b*) with better power behavior for the typical one-sided hypothesis: multiple end point analysis (*49*) based on Dunnett's procedure (*50*) is a special case of parametric testing after *k*-ranking transformation. Both approaches have, however, a major disadvantage: only decision of the global end point vector. Information is not available on the combinations of end points, which might go as far as the single end point case.

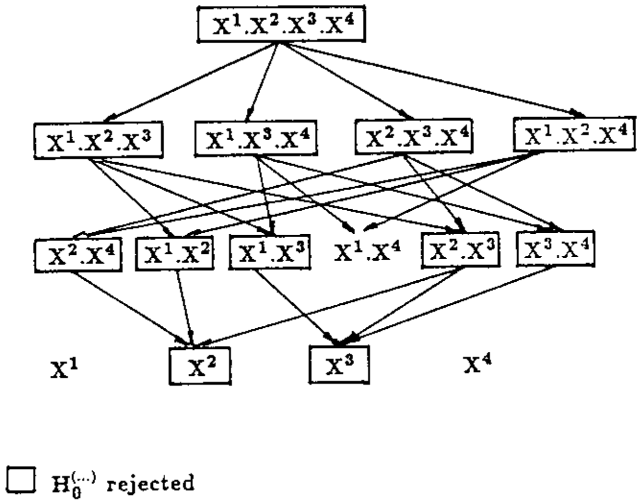

FIGURE 2. Complete hypotheses system in the case of four end points.

The multivariate problem also consists of a complete hypothesis system of $2^k - 1$ elementary hypotheses (*21*). The decision scheme is quite simple (*51*), as can be seen for the four end points in Figure 2. Based on the level α-test on each step, this procedure shows good power behavior. This procedure is available as a PC program for up to 10 end points (Hothorn and Nagel, submitted).

An interesting extension of this method is possible for toxicity studies with both multiple end points and multiple treatment or dose groups based on the closed testing procedure under the assumption of an ordered alternative using Williams (*17*) MCP. With this approach, decisions can be performed both on the multiple end points and the multiple dose group based on level α tests on each step, but holding α_{exp} (*50*).

Summary

This paper reveals several sources of multiplicity within long-term toxicity studies and their suitable treatment, the possibility of reducing the antagonism between holding α_{exp} and ensuring the maximum power, that special MCPs for biostatistical analysis of long-term toxicological studies are necessary and are available as a PC program.

REFERENCES

1. Marcus, R., Peritz, E., and Gabriel, K. R. On closed testing procedure with special reference to ordered analysis of variance. Biometrika 63:655–660 (1976).
2. Anonymous. Empfehlungen des rates vom 26.10.1983 zu den Versuchen mit Arzneimittelspezialitaeten im Hinblick auf deren Inverkehrbringung (EWG 83/571). Pharmazeut. Ind. 45:1248–1261 (1983).
3. Margolin, B. H., Kaplan, N., and Zeiger, E. Statistical analysis of the Ames *Salmonella*/microsome test. Proc. Natl. Acad. Sci. U.S.A. 78: 3779–3783 (1981).
4. Jossan, S. S. MPTP toxicity in relation to age, dopamine uptake and MOA-B activity in two rodent species. Pharmacol. Toxicol. 64: 314–318 (1989).
5. Hothorn, L. General principles in testing of toxicological studies. In: Statistical methods in toxicology. Lecture Notes in Medical Informatics, vol. 43 (Hothorn, L., Ed.), Sprinter-Verlag, Heidelberg, 1991, pp. 111–131.
6. Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. J. Am Stat. Assoc. 50: 1096–1121 (1955).
7. Piegorsch, W. W. Multiple comparisons for analyzing dichotomous response. Biometrics 47: 45–52 (1991).
8. Dunnett, C. W. New tables for multiple comparisons with a control. Biometrics 20: 560–572 (1964).
9. Gupta, S. S. On the distribution of the studentized maximum of equally correlated normal variables. Commun. Stat. B14: 103–135 (1985).
10. Bechhofer, R. E., and Dunnett, C. W. Tables of percentage points of multivariate t distributions. In: Selected Tables in Mathematical Statistics, No. 11. American Mathematics Society, Providence, RI, 1988, pp. 1-112.
11. Ahner, C., and Passing, H. Berechnung der multivariaten t-Verteilung and simultane Vergleiche gegen Kontrolle bei ungleichen Gruppenbesetzungen. EDV Med. Biol. 14: 113-120. (1983).
12. Rudolph, P.E. Robustness of multiple comparison procedures: treatment versus control. Biometrics J. 30: 41-45 (1988).
13. Steel, R. G. D. A multiple comparison rank sum test treatment versus control. Biometrika 15:560–572 (1959).
14. Horn, M. Regarding the optimality of the formula $n_0/n = \sqrt{p}$ for the ratio of sample sizes of a control and p treated groups. Biometrics J. 21: 407–412 (1979).
15. Hochberg, Y., and Tamhane, A. C. Multiple Comparison Procedures. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1987.
16. Spurrier, J. D., and Nizam, A. Sample size allocation for simultaneous inference in comparison with control experiments. J. Am. Stat. Assoc. 85: 181-186 (1990).
17. Williams, D. A. A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics 27: 103–117 (1971).
18. Hothorn, L. Robustness study on Williams- and Shirley- procedure, with application in toxicology. Biometrics J. 31: 891–903 (1989).
19. Hothorn, L. Biometrische Analyse spezieller Untersuchungen der regulatorischen Toxikologie. In: Aktuelle Probleme der Toxikologie, Vol. 5 Grundlagen der Statistik fuer Toxikologen (M. Horn and L. Hothorn, Eds. Verlag Gesundheit Gmbh, Berlin, 1990, pp. 130–236.
20. Brownie, C., Boos, D. D., and Hughes-Oliver, J. Modifying the t and ANOVA F test when treatment is expected to increase variability relative to controls. Biometrics 46: 259–266 (1990).
21. Sonnemann, E. Allgemeine Loesung multiplier Testprobleme. EDV Med. Biol. 3: 120–128 (1982).
22. Holm, S. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6: 65–70 (1979).
23. Hayter, A. J., and Tamhane, A. C. Sample size determination for step-down multiple test procedures: orthogonal contrasts and comparison with a control. J. Stat. Plan. Infer. 27: 271–290 (1991).
24. Hommel, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75: 383–386 (1988).
25. Hommel, G. A comparison of two modified Bonferroni procedures. Biometrika 76: 624–625 (1989).
26. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75: 800–802 (1988).
27. Rom, D. A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika 77: 663–665 (1990).
28. Hothorn, L., and Lehmacher, W. A simple testing procedure 'control versis k treatments' for one-sided ordered alternatives, with application in toxicology. Biometrics J. 33: 179–189 (1991).
29. Fligner, M. A., and Wolfe, D. A. Distribution-free tests for comparing several treatments with a control. Stat. Neerl. 36: 119–127 (1982).
30. Williams, D. A. The comparison of several dose levels with a zero dose control. Biometrics 28: 519–531 (1972).
31. Shirley, E. A. C. A non-parametric equivalent of Williams test for contrasting inceasing dose levels of a treatment. Biometrics 33: 386–389 (1977).
32. Williams, D. A. A note on Shirley's non-parametric test for comparing several dose levels with a zero-dose control. Biometrics 42: 183–186 (1986).
33. Armitage, P. Tests for linear trends in proportions and frequencies. Biometrics 11: 375–386 (1955).
34. Lee, Y. J. Tests in trends in Poisson means. J. Qual. Technol. 17: 44–49 (1985).
35. Shirley, E. A. C. The comparison of treatment with control group means in toxicological studies. Appl. Stat. 28: 144–151. (1979).
36. Mukerjee, H., Robertson, T., and Wright, F. T. Comparison of several treatments with a control using multiple contrasts. J. Am Stat. Assoc. 82: 902–910 (1987).

37. Ruberg, S. J. Contrasts for identifying the minimum effective dose. J. Am. Stat. Assoc. 84:P 816–822 (1989).
38. Ansari, A. R., and Bradley, R. A. Rank-sum tests for dispersions. Ann. Math. Stat. 31: 1174–1189 (1960).
39. Lepage, Y. A combination of Wilcoxon's and Ansari-Bradley's statistics. Biometrika 58: 213–217 (1971).
40. Lehmann, E. L. The power of rank tests. Ann. Math. Stat. 24: 23–43 (1953).
41. Good, P. I. Detection of a treatment effect when not all experimental subjects will respond to treatment. Biometrics 35: 483–489 (1979).
42. Johnson. R. A., Verrill, S., and Moore, D. H. Two-sample rank tests for detecting changes that occur in a small proportion of the treated population. Biometrics 43: 641–655 (1987).
43. Conover, W. J., and Salsburg, D. S. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. Biometrics 44: 189–196 (1988).
44. Nation, J. R. The effect of oral cadmium exposure on a passive avoidance performance. Toxicol. Let. 20: 41–47 (1984).
45. Cory-Slechta, D. A. Chronic postweaning lead exposure and response duration performance. Toxicol. Appl. Pharmacol. 60: 78–84 (1981).
46. Hothorn, L. Biostatistical analysis of the micronucleus mutagenicity assay based on the assumption of a mixing distribution. Environ. Health Perspect 102(Suppl 1):121–125 (1994).
47. Unkelbach, H.-D., Deyssenroth, G., Helmstaedter, G., Knappen, F., Luedin, E., Mau, J., Passing, H., and Peil, H. Statische Auswertung haematoligischer and klinisch-chemischer Daten: Derzeitiger Stand bei toxikologischen Standardversuchen. In: Biometrie in der chemisch-pharmazeutischen Industrie, Vol. 1 (J. Vollmar, Ed.), G. Fischer Verlag, Stuttgart, 1983, pp. 45–56.
48. Higazi, S. M. F., and Dayton, C. M. Comparing several experimental groups with a control in the multivariate case. Commun. Stat. B13: 227–241 (1984).
49. O'Brien, P. C. Procedures for comparing samples with multiple endpoints. Biometrics 40: 1079–1087 (1984).
50. Hothorn, L. Multivariate testing in toxicological studies for *control versus k treatment or dose groups* design. In: Proceedings of the 15th International Biometrics Conference, Budapest, 1990, p. 102.
51. Lehmacher, W., Wassmer, G., and Reitmeir, P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics 47: 511–522 (1991).